EAST Search History

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S1	0	"20010317818" [pn]	US-PGPUB; USPAT	AND	ON	2007/07/09 15:37
S2	0	"2001317818" [pn]	US-PGPUB; USPAT	AND	ON	2007/07/09 15:34
S3	0	"20010317824" [pn]	US-PGPUB; USPAT	AND	ON	2007/07/09 15:35
S4	1	8001939CD1	US-PGPUB; USPAT	AND	ON	2007/07/09 15:34
S5	1	"0317824" [pn]	US-PGPUB; USPAT	AND	ON	2007/07/09 15:36
S6	44419	Tang	US-PGPUB; USPAT	AND	ON	2007/07/09 15:36
S7	585	S6 and Incyte	US-PGPUB; USPAT	AND	ON	2007/07/09 15:36
S8	522	S7 and immunoglobulin	US-PGPUB; USPAT	AND	ON	2007/07/09 15:36
S9	0	S8 nad "0317818"	US-PGPUB; USPAT	AND	ON	2007/07/09 15:36
S10	1	S8 and "8001939"	US-PGPUB; USPAT	AND	ON	2007/07/09 15:37
S11	0	"2001-0317818" [pn]	US-PGPUB; USPAT	AND	ON	2007/07/09 15:37
S12	0	"010317818" [pn]	US-PGPUB; USPAT	AND	ON	2007/07/09 15:38
S13	4	"2003016506" [pn]	EPO; JPO; DERWENT	AND	ON	2007/07/09 15:39

RESULT 1
US-10-487-078-25
; Sequence 25, Application US/10487078
; Publication No. US20050064543A1
; GENERAL INFORMATION:
; APPLICANT: INCYTE CORPORATION; TANG et al
; TITLE OF INVENTION: SECRETED PROTEINS
; FILE REFERENCE: PF-1141 USN
; CURRENT APPLICATION NUMBER: US/10/487,078
; CURRENT FILING DATE: 2004-02-18
; PRIOR APPLICATION NUMBER: PCT/US02/27143
; PRIOR FILING DATE: 2002-08-15
; PRIOR APPLICATION NUMBER: US 60/313,249
; PRIOR FILING DATE: 2001-08-17
; PRIOR APPLICATION NUMBER: US 60/314,752
; PRIOR FILING DATE: 2001-08-24
; PRIOR APPLICATION NUMBER: US 60/317,818
; PRIOR FILING DATE: 2001-09-07
; PRIOR APPLICATION NUMBER: US 60/317,824
; PRIOR FILING DATE: 2001-09-07
; PRIOR APPLICATION NUMBER: US 60/324,040
; PRIOR FILING DATE: 2001-09-21
; PRIOR APPLICATION NUMBER: US 60/324,586
; PRIOR FILING DATE: 2001-09-24
; PRIOR APPLICATION NUMBER: US 60/343,980
; PRIOR FILING DATE: 2001-11-02
; PRIOR APPLICATION NUMBER: US 60/334,229
; PRIOR FILING DATE: 2001-11-28
; PRIOR APPLICATION NUMBER: US 60/357,002
; PRIOR FILING DATE: 2002-02-13

See also
PN WO2003016506-A2.
XX
PD 27-FEB-2003.
XX
PF 15-AUG-2002; 2002WO-US027143.
PR 17-AUG-2001; 2001US-0313249P.
PR 24-AUG-2001; 2001US-0314752P.
PR 07-SEP-2001; 2001US-0317818P.
PR 07-SEP-2001; 2001US-0317824P.
PR 21-SEP-2001; 2001US-0324040P.
PR 24-SEP-2001; 2001US-0324586P.
PR 02-NOV-2001; 2001US-0343980P.
PR 28-NOV-2001; 2001US-0334229P.
PR 13-FEB-2002; 2002US-0357002P.
PR 06-MAR-2002; 2002US-0362439P.
PR 19-MAR-2002; 2002US-0366041P.
PR 30-APR-2002; 2002US-0376988P.

; Remaining Prior Application data removed - See File Wrapper or PALM.
; SEQ ID NO 25

; LENGTH: 270
; TYPE: PRT
; ORGANISM: Homo sapiens
; FEATURE:
; NAME/KEY: misc_feature
; OTHER INFORMATION: Incyte ID No: 8001939CD1
US-10-487-078-25

Query Match 100.0%; Score 1450; DB 5; Length 270;
Best Local Similarity 100.0%; Pred. No. 7.4e-121;
Matches 270; Conservative 0; Mismatches 0; Indels 0; Gaps 0;

```
Qy          1 MENQPVRWRALPGLPRPPGLPAAPWLLLGVLLLPGTLRLAGGQSVTHTGLPIMASLANTA 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db          1 MENQPVRWRALPGLPRPPGLPAAPWLLLGVLLLPGTLRLAGGQSVTHTGLPIMASLANTA 60

Qy         61 ISFSCRITYPYTPQFKVFTVSYFHEDLQGQRSPKKPTNCHPGLGTENQSHTLDCQVTLVL 120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db         61 ISFSCRITYPYTPQFKVFTVSYFHEDLQGQRSPKKPTNCHPGLGTENQSHTLDCQVTLVL 120

Qy        121 PGASATGTYYCSVHWPHSTVRGSGTFILVRDAGYREPPQSPQKLLLFGFTGLLSVLSVVG 180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db        121 PGASATGTYYCSVHWPHSTVRGSGTFILVRDAGYREPPQSPQKLLLFGFTGLLSVLSVVG 180

Qy        181 TALLLWNKKRMRGPGKDPTRKCPDPRSASSPKQHPSESVYTALQRRETEVYACIENEDGS 240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db        181 TALLLWNKKRMRGPGKDPTRKCPDPRSASSPKQHPSESVYTALQRRETEVYACIENEDGS 240

Qy        241 SPTAKQSPLSQERPHRFEDDGELNLVYENL 270
              |||||||||||||||||||||||||||||
Db        241 SPTAKQSPLSQERPHRFEDDGELNLVYENL 270
```

Deciphering the Message in Protein Sequenc Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM, ROBERT T. SAUER

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (*1*), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (*2–7*). The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or sc identify functional sequences. This approach has been used advantage for proteins that can be expressed in bacteria o where the appropriate genetic manipulations are possible (*3* The end results of both methods are lists of active sequences be compared and analyzed to identify sequence features essential for folding or function. If a particular property o chain, such as charge or size, is important at a given positi side chains that have the required property will be allowe versely, if the chemical identity of the side chain is unim then many different substitutions will be permitted.

Studies in which these methods were used have reveal proteins are surprisingly tolerant of amino acid substitutio *11*). For example, in studying the effects of approximatel single amino acid substitutions at 142 positions in *lac* re Miller and co-workers found that about one-half of all subst were phenotypically silent (*11*). At some positions, many d nonconservative substitutions were allowed. Such residue p play little or no role in structure and function. At other posit substitutions or only conservative substitutions were allowe residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains proteins? Residues that are directly involved in protein fi such as binding or catalysis will certainly be among th conserved. For example, replacing the Asp in the catalytic trypsin with Asn results in a 10^4-fold reduction in activity similar loss of activity occurs in λ repressor when a DNA residue is changed from Asn to Asp (*13*). To carry o function, however, these catalytic residues and binding must be precisely oriented in three dimensions. Conse mutations in residues that are required for structure form stability can also have dramatic effects on activity (*10*, Hence, many of the residues that are conserved in sets of sequences play structural roles.

Substitutions at Surface and Buried Posit

In their initial comparisons of the globin sequences, Per co-workers found that most buried residues require nonp chains, whereas few features of surface side chains are g conserved (*6*). Similar results have been seen for a number of families (*2*, *4*, *5*, *7*, *17*, *18*). An example of the sequence tole surface versus buried sites can be seen in Fig. 1, which sh allowed substitutions in λ repressor at residue positions that

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

Bioscience Reports, Vol. 8, No. 4, 1988

Several Epitopes of p85 Glycoprotein (CDw44) are Dependent on Intact Disulphide Bonds. Isolation of cDNA Clones Requires a Polyclonal Antibody Raised Against the Reduced Protein

Ian Rogers,[1] Giacomo D'Agostaro,[2] Sonia Vera[1] and Michelle Letarte[1,3]

Received April 22, 1988

Monoclonal antibodies 50B4 and 50E6 recognize two distinct epitopes of human p85 glycoprotein (CDw44). Both epitopes are destroyed by reduction of the purified glycoprotein as demonstrated by inhibition of cellular radioimmunoassay and Western blot analysis. Endoglycosidase F treated p85 glycoprotein, with an apparent molecular weight of 73,000, is still reactive with both monoclonal antibodies. Thus both epitopes are conformational determinants of the polypeptide chain. A rabbit antibody produced against purified native p85 glycoprotein also reacted only with the non-reduced form of p85. Repeated immunizations with SDS-dissociated and reduced p85 yielded a polyclonal antibody reactive by Western blot analysis with reduced and non-reduced forms of p85 glycoprotein. When a HOON leukemia cell line cDNA expression library was screened with this polyclonal antibody, two cDNA clones were isolated which reacted specifically with the antiserum and not with the control non-immune serum. Preliminary characterization of these clones indicates that they are p85-related.

KEY WORDS: CDw44; conformational epitopes; lymphocyte antigen.

INTRODUCTION

Human p85 glycoprotein was first identified with MAb F10-44-2 as a brain-leukocyte antigen (1, 2). This antibody included in the Third International Workshop on human leukocyte differentiation antigens defined a new cluster, CDw44 (3). Several MAb to the p85 glycoprotein were obtained by immunization

Departments of Immunology[1] and Medical Genetics,[2] University of Toronto and [1]Division of Immunology, the Hospital for Sick Children, 555 University Avenue, Toronto, M5G 1X8
[2] Current address: Lab. Biophysics, CRE Enea, S.P. Anguillarese, 301 Casaccia, 00100 Roma.
[3] To whom correspondence should be addressed.

0144-8463/88/0800-0359$06.00/0

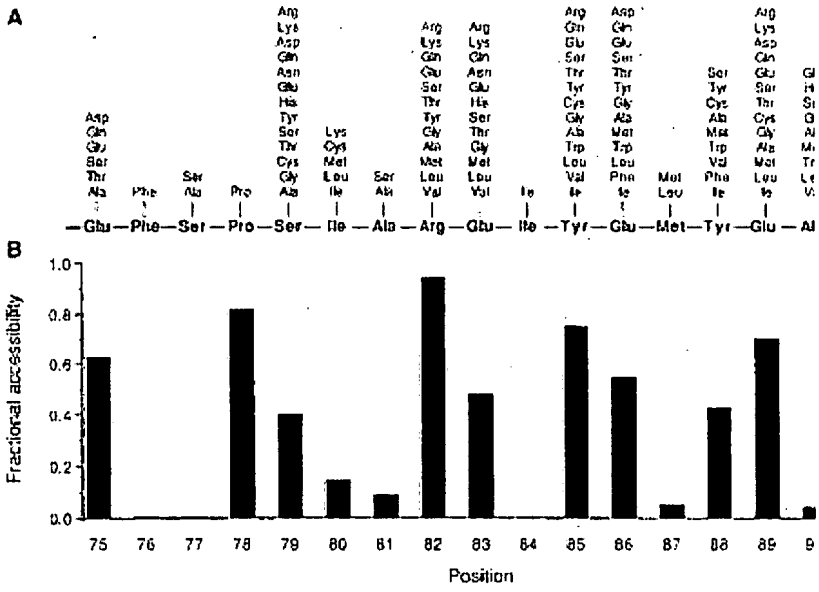

Fig. 1. (**A**) Amino acid substitutions allowed in a short region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (*9*). (**B**) The fractional solvent accessibility (*42*) of the wild-type side chain in the protein dimer (*43*) relative to the same atoms in an Ala-X-Ala model tripeptide.

selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (*19*). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (*20*). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (*21*). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (*22*).

The cores of most proteins are quite closely packed (*23*), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core

phylogenetic studies, where it has been noted that the size c and increases at interacting residues are not necessarily rel simple complementary fashion (*5, 7, 17*). Rather, local changes are accommodated by conformational changes i side chains and by a variety of backbone movements.

The Informational Importance of the Co

With occasional exceptions, the core must remain hydr and maintain a reasonable packing density. However, since is composed of side chains that can assume only a limited n conformations (*24*), efficient packing must be maintained steric clashes. How important are hydrophobicity, volu steric complementarity in determining whether a given sequ form an acceptable core? Each factor is essential in a physic as a stable core is probably unable to tolerate unsatisfied h bonding groups, large holes, or steric overlaps (*25*). Howev informational sense, these factors are not equivalent. For ex experiments in which three core residues of λ repress mutated simultaneously, volume was a relatively uninporta mational constraint because three-quarters of all possible c tions of the 20 naturally occurring amino acids had volume the range tolerated in the core, and yet most of these sequen unacceptable (*20*). In contrast, of the sequences that contai

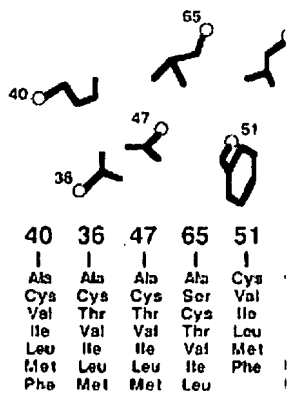

Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (*43*). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (*20*). Not all substitutions are al-

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (*26*), or by binding to antibodies specific for the native structure (*27, 28*). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (*8*). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (*28*). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 26, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33, 34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4, 29, 38, 39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple tion is shown in Fig. 5. Residues adjacent in the sequen occupy adjacent squares on the lattice, and two residues occupy the same space. Free energies of particular conforma evaluated with a single term, an attraction of H gro considering chains of ten residues, an exhaustive confor search for all 1024 possible sequences of H and P resid possible. For longer sequences only a representative fractio allowed sequence or conformation space could be explor significant results were as follows: (i) not all sequences can f a "native" structure and only a few sequences form a uniqu structure; (ii) the probability that a sequence will adopt a native structure increases with chain length; and (iii) th states are compact, contain a hydrophobic core surrounded residues, and contain significant secondary structure. Altho gap between these two-dimensional simulations and three sional structures is large, the use of simple rules and s representations yields results similar to those expected proteins. Three-dimensional lattice methods are also begir be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences t single sequence. A number of practical applications arise analysis of the tolerance of residue positions to change. Fi information permits the evaluation of a residue's importanc function and stability of a protein. This ability to iden essential elements of a protein sequence may improve our standing of the determinants of protein folding and stability as protein function. Second, patterns of tolerance to ami substitutions of varying hydrophilicity can help to identify likely to be buried in a protein structure and those likely to

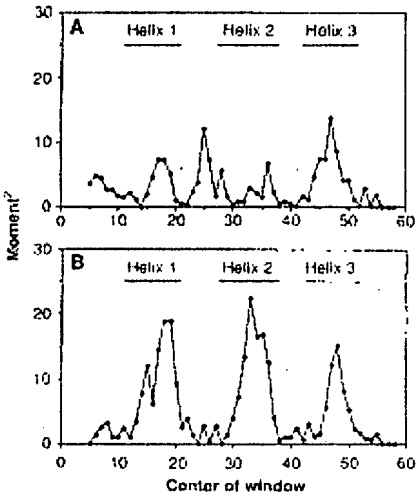

Fig. 4. Helica phobic momen lated by using Antennapedia h main sequence set of 39 aligne domain sequen The bars indica rent of the h gions identified ar magnetic r studies of the pedia home (*36*). To deter drophobic n residues were to one of three H1 (high hydr ty = Trp, Ile, P Met, Val, or C (medium hydr ity = Tyr, Pro, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, / or Arg). For the aligned homeodomain sequences, the residues position were sorted by their hydrophobicity by using the scale of I and Pliska (*45*). Arg and Lys were not counted unless no other res found at the position, because they contain long aliphatic side chain thereby substitute for nonpolar residues at some buried sites. To acc possible sequence errors and rare exceptions, the most hydrophili allowed at each position was discarded unless it was observed tw second most hydrophilic residue was then chosen to represent the hy bicity of each position. An eight-residue window was used and th

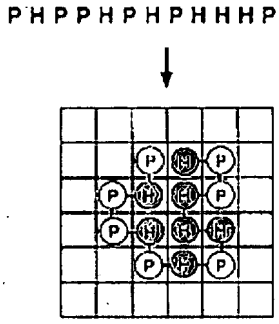

Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (40), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, P. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. I. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (5) have argued that a protein core composed solely of buried residues would also be inviable on evolutionary grounds, as a change in one core residue would require compensating changes in any residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baldwin, E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); L. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Bioc T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); I. Greer, *ibid.* **153**, 102 M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim *Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.* (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9 M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. P 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Fun* preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambri
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem. Chim. Ther.* **18**, 36
46. We thank C. O. Pabo and S. Jordan for coordinates of the NH₂-termin λ repressor and its operator complex. We also thank P. Schimmel for graphics system and J. Bumbaum and C. Francklyn for assistance. S part by NIH grant AI-15706 and predoctoral grants from NSF (J Howard Hughes Medical Institute (W.A.L.).